Wang et al. BMC Genomics 2011, 12(Suppl 5):S8
http://www.biomedcentral.com/1471-2164/12/S5/S8

BMC Genomics

RESEARCH Open Access

Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1

Xin Wang[1,2], Liran Juan[1,3], Junjie Lv[2], Kejun Wang[2], Jeremy R Sanford[4], Yunlong Liu[1,3,5*]

From BIOCOMP 2010. The 2010 International Conference on Bioinformatics and Computational Biology Las Vegas, NV, USA. 12-15 July 2010

Abstract

Background: RNA-binding proteins (RBPs) play diverse roles in eukaryotic RNA processing. Despite their pervasive functions in coding and noncoding RNA biogenesis and regulation, elucidating the sequence specificities that define protein-RNA interactions remains a major challenge. Recently, CLIP-seq (Cross-linking immunoprecipitation followed by high-throughput sequencing) has been successfully implemented to study the transcriptome-wide binding patterns of SRSF1, PTBP1, NOVA and fox2 proteins. These studies either adopted traditional methods like Multiple EM for Motif Elicitation (MEME) to discover the sequence consensus of RBP's binding sites or used Z-score statistics to search for the overrepresented nucleotides of a certain size. We argue that most of these methods are not well-suited for RNA motif identification, as they are unable to incorporate the RNA structural context of protein-RNA interactions, which may affect to binding specificity. Here, we describe a novel model-based approach–*RNAMotifModeler* to identify the consensus of protein-RNA binding regions by integrating sequence features and RNA secondary structures.

Results: As an example, we implemented *RNAMotifModeler* on SRSF1 (SF2/ASF) CLIP-seq data. The sequence-structural consensus we identified is a purine-rich octamer 'AGAAGAAG' in a highly single-stranded RNA context. The unpaired probabilities, the probabilities of not forming pairs, are significantly higher than negative controls and the flanking sequence surrounding the binding site, indicating that SRSF1 proteins tend to bind on single-stranded RNA. Further statistical evaluations revealed that the second and fifth bases of SRSF1octamer motif have much stronger sequence specificities, but weaker single-strandedness, while the third, fourth, sixth and seventh bases are far more likely to be single-stranded, but have more degenerate sequence specificities. Therefore, we hypothesize that nucleotide specificity and secondary structure play complementary roles during binding site recognition by SRSF1.

Conclusion: In this study, we presented a computational model to predict the sequence consensus and optimal RNA secondary structure for protein-RNA binding regions. The successful implementation on SRSF1 CLIP-seq data demonstrates great potential to improve our understanding on the binding specificity of RNA binding proteins.

Introduction

RNA-binding proteins (RBPs) are implicated in virtually every step of post-transcriptional gene expression including pre-mRNA splicing, RNA editing and polyadenylation [1]. These proteins possess a diverse array of structurally and functionally distinct RNA-binding domains such as RNA recognition motifs (RRM), KH domains, RGG boxes, zinc finger, double-stranded RNA-binding domain, etc [1]. In contrast to DNA, recognition sites for RNA binding proteins can be presented diverse structural contexts. Indeed the structural context of binding sites can have pronounced effects on protein-RNA interactions [2,3]. Likewise, RNA binding proteins can alter the folding landscape of RNA

* Correspondence: yunliu@iupui.edu
[1]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, IN 46202, USA
Full list of author information is available at the end of the article

molecules thereby inducing structured or single stranded conformations [4]. Given the significant role RNA folding plays in promoting or inhibiting protein-RNA interactions methods for evaluating both the sequence and RNA-structural determinants to binding specificity will be highly beneficial to the field.

Several methods for elucidating the specificity of protein-RNA interactions enable rapid advances in our understanding of RBP functions. One recent innovation is the Cross-Linking ImmunoPrecipitation (CLIP); CLIP exploits photoreactive residues in RNA and polypeptides to generate covalently linked complexes. Because UV irradiation does not induce protein-protein cross-links CLIP is thought to be more specific than other IP based assays for protein-RNA interactions. CLIP was successfully applied to identify mRNA targets of the NOVA protein, a neural splicing factor associated with paraneoplasticopsoclonus myoclonus ataxia (POMA) [5-7]. Coupling CLIP with next-generation high-throughput sequencing technology, known as CLIP-seq or HITS-CLIP, provides a cost-efficient method to increase the sensitivity of the assay by surveying the RNA landscape on a more global scale. Several groups have successfully implemented CLIP-seq analysis of NOVA, SRSF1, fox2 and PTB proteins in mammalian systems [5,8-10]. Both MEME and Z-score statistics have been used to reveal consensus binding motifs that are overrepresented in CLIP-seq data [5,9]. Although Z-score statistics may be able to identify the overrepresented sequence motifs, it does not consider the degenerate feature of the binding specificities of RBPs. MEME-based method is well known to be an excellent tool for cases only regarding sequence specificity [11]. Neither of these approaches canascertain the roles of RNA secondary structure in establishing the context of the protein-RNA interaction. Hiller et al. extended MEME by adding a pre-computing procedure to measure single-strandedness of RNA sequence as *a prior* knowledge to guide the motif search. They demonstrated that their model, MEMERIS, is able to identify binding motifs located in single-stranded regions with both artificial and biological data [12]. Recently, Kazan et al. proposed *RNAcontext* for learning both sequence and structural binding preferences of RNA-binding proteins [13].

Here we describe a model-based approach–*RNAMotifModeler* to evaluate protein-RNA interactionsusinga retained binding affinity ratio, which is considered to be affected by two major factors–sequence degeneracy and RNA secondary structure deviation. *RNAMotifModeler* incorporates predicted unpaired probability of each nucleotide in the protein-RNA binding regions; such probability is derived from RNA secondary prediction algorithms, such as RNA-fold, based on the nucleotide compositions of the neighbouring flanking sequences. This strategy is different from *RNAContext*, which uses predicted RNA secondary structures as input such as 'Paired', 'Hairpin Loop', 'Unstructured' or 'Miscellaneous'. Unlike *MEMERIS*, *RNAMotifModeler* uses the base-pairing probability for each nucleotide rather than the entire binding site. For each binding instance, *RNAMotifModeler* defines a score that evaluates the consensus binding site within an optimal structural context, and aims at searching for an optimal RNA sequence-structural consensus for an RNA binding protein. These features enhance our ability to calculate and estimate the sequences that yield the highest binding affinity for a specific RBP.

We tested *RNAMotifModeler* on CLIP-seq data that profile the transcriptome-wide binding pattern of SRSF1, serine/arginine-rich splicing factor 1 [5]. The sequence features of the binding motifs are consistent with the experimentally defined *cis*-acting elements recognized by SRSF1 [5,14,15]. Interestingly, the prediction suggests that the second and fifth bases of SRSF1octamer motif have stronger sequence specificities, but lower p-values of unpaired probabilities, while the third, fourth, sixth and seventh bases are more significantly to be single-stranded, but have less sequence specificities. Therefore, we conclude that the sequence and structure specificities are both required and may play complementary roles during binding site recognition of SRSF1.

Results

Elucidating the sequence and structural features defining protein-RNA interactions is a major challenge in the field. To begin to address this problem we developed a tool to evaluate the structural context of RNA fragments co-purified with RNA binding proteins by CLIP. The results presented here focus on SRSF1; however this tool will be generally applicable to any RNA binding protein. SRSF1 is an essential splicing factor with multiple roles in post-transcriptional gene expression [16]. SRSF1 is also a potent proto-oncogene and implicated in maintaining genome stability [17]. Moreover, loss of SRSF1 binding sites by mutations linked to genetic diseases can induce aberrant patterns of pre-mRNA splicing [5]. Thus considerable effort has been focused on defining the binding specificity and RNA targets of SRSF1. Here we report a novel tool intended to examine the contributions of structural and sequence elements in RNA fragments co-purified with SRSF1 by CLIP.

Workflow of RNAMotifModeler

The first step of RNAMotifModeler is to do data preprocessing. In this study, the data came from our previous genome-wide profiling of SRSF1 protein's binding sites by combining cross-linking immunoprecipitation

(CLIP) with high-throughput sequencing [5]. In total, 932,152 reads were obtained from SFRS1-bound RNA in four independent experiments. As a comparison, 670,448 reads were generated from three experiments performed on nonselected input RNA. After removing redundant sequences and alignment to the human genome, we obtained 953 and3374 loci for CLIP and input RNA samples, respectively.904 positive gold standard sequences were selected from at least three out of the four CLIP-seq experiments and absent from the input sequences. A same number of negative sequences were randomly picked from non-SRSF1-targeted regions belonging to the same genomic category (exonic, intronic, intergenic, etc) as their positive counterparts. Base-pairing probabilities of each nucleotide to its neighbours were subsequently predicted by *RNAfold* [18] (ViennaRNA package, version 1.7.2) for both positive and negative gold standard sequences.

Our next step, as shown in Figure 1, is to identify sequence-structural consensus for the RNA binding protein based on the gold standard sequences from CLIP-seq data. We took an iterative approach that alternates between: 1) optimization of parameters specifying sequence degeneracy and structural context given a sequence motif, and 2) search of optimal sequence reference motif given the estimated parameters by evaluation of each motif candidate's contribution to binding affinities of positive gold standard sequences. The above two steps will be repeated until a convergence when the starting motif candidate makes the most contribution to binding affinities.

Finally, RNAMotifModeler outputs the converged sequence motif, optimal parameters, statistical performance of using the optimal parameters such as the area under the ROC curve (AUC), etc. The AUC scores are measured by area under ROC curves derived from predictions of gold standard sequences being bound by SRSF1 proteins varying the binding affinity threshold. In order to predict binding sites of SRSF1 proteins, we pick the sequence binding affinity yielding the maximal prediction accuracy as a cutoff score. Based on the predicted parameters, positive gold-standard sequences can be scanned to find all potential binding sites with binding affinities higher than the cutoff score. These binding sites can be further used for sequence logo creation and transformed to positional weight matrix, which aremuch more widely used.

Convergence of SRSF1 consensus motif searching

We call the converging path from a starting motif candidate to the final consensus motif a *motif searching pathway*. To have a global overview of the convergence, motif searching pathways for all motif candidates are organised together to form a *motif searching graph*. In

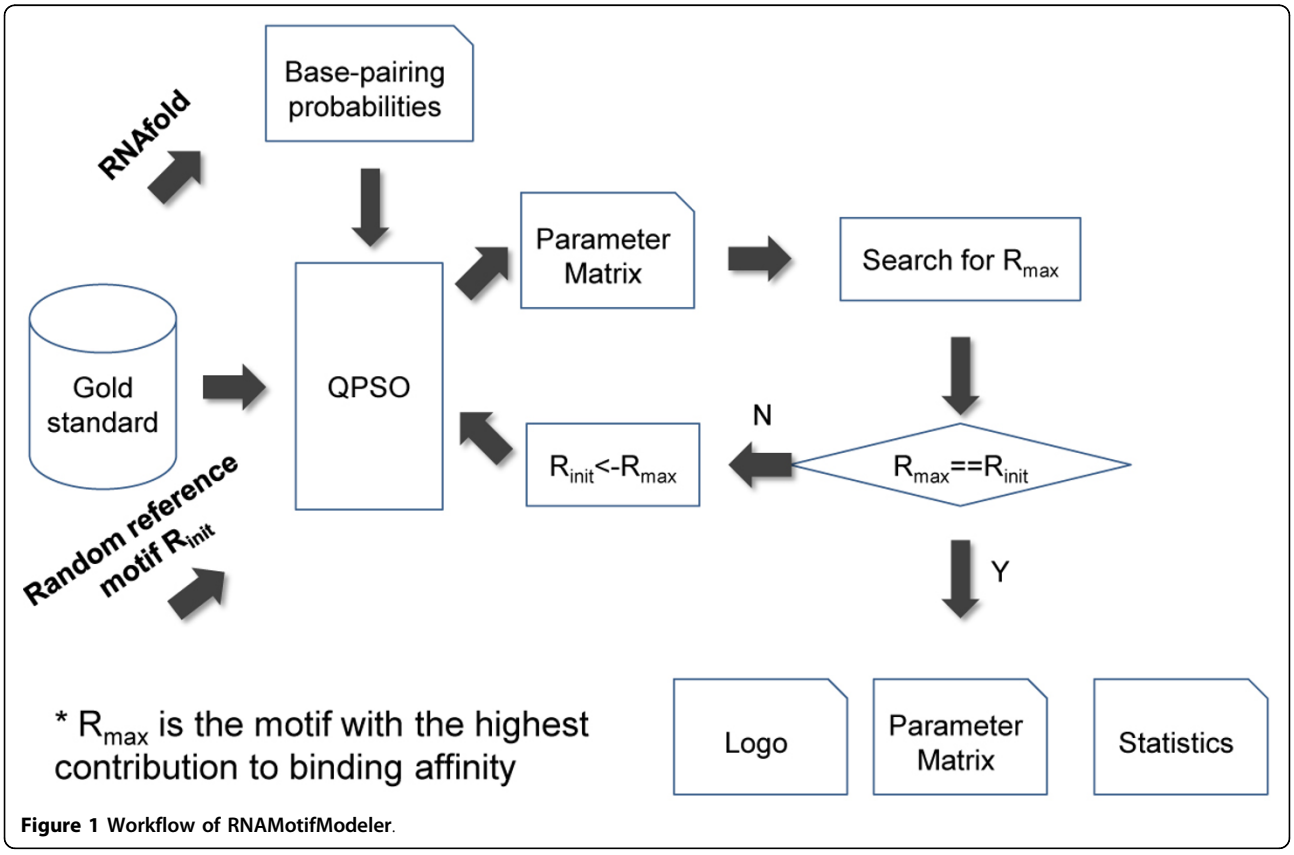

Figure 1 Workflow of RNAMotifModeler.

the particular case of hexamer prediction for SRSF1, the motif searching pathwaysof all initialized reference motifs converge to a short list of candidates (Figure 2). All of the 4096 motif candidates converge within three iterations, of which 85.7% converge after the first iteration. AGAAGA, AAGAAG and GAAGAA are top three hexamers with the highest in-degrees, responsible for 99.7% of all motif candidates (Table 1). Despite only one or two sequence alterations, the other twelvereference motifs are closely related to these three motifs. It is also notedthat nearly an equal number of motif candidates convergeto each one of the top three reference motifs. More interestingly, these hexamers share a core sequence of 'AAGA'indicating that they may be adjacent to each other in RNA fragments.

Table 1 The final converged motifs and their corresponding numbers of source motifs

Converged motif	No. of source motifs
AGAAGA	1484
AAGAAG	1375
GAAGAA	1225
Others	12

SRSF1 consensus motifs of different lengths

RNAMotifModeler provides an option to predict sequence-structural consensus of different lengths. We have mentioned in the previous section that for short motifs, it is suggested to perform predictions starting from every potential motif candidate and generate a motif searching graph to inspect the global convergence. For longer motifs, however, it is computationally expensive. In this case, we conduct predictions starting froma sufficient number of motif candidates randomly picked from motif space. The converged motif with the highest prediction power, measured by AUC, is selected as the optimal one.

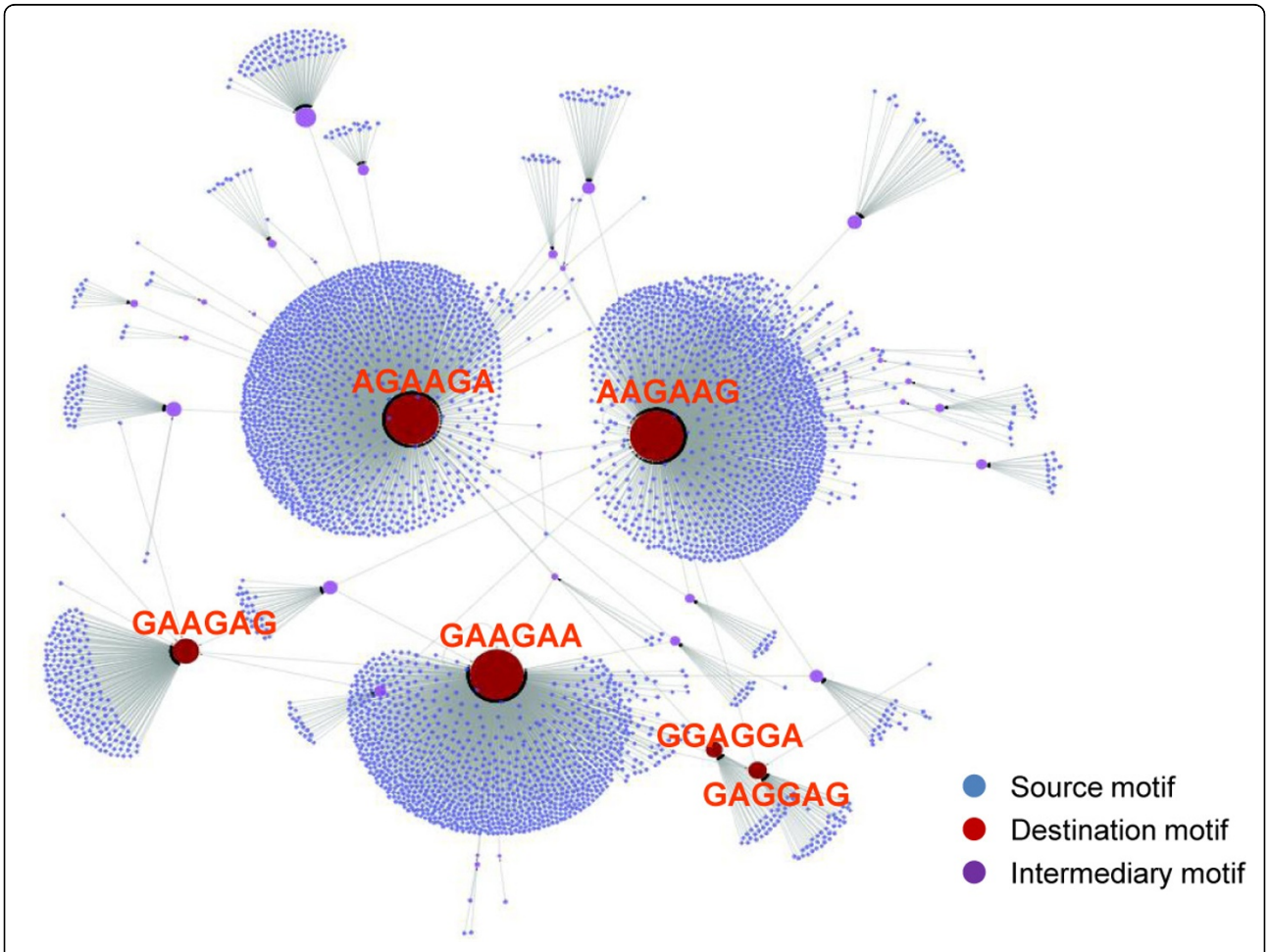

Figure 2 Motif searching graph. Source, intermediary and destination motifs are represented by nodes colored in blue, purple and red. The size of node is proportional to its in-degree. Arrows between nodes indicate reference motif searching pathways. This figure demonstrates the fast convergence of the vast majority of motif candidates using the Quantum Particle Swarm Optimization algorithm.

Using the above strategy, we predicted optimal 6nt, 7nt and 8nt consensus motifs for SRSF1 proteins (Additional file 1(A), (B) and Table 2). Interestingly, the sequence motifs of different lengths are highly similar to each other. Comparing their sequence and structural parameters identified, we can also see a high consistency among them. Importantly, the predicted unpaired probabilities of these three motifs indicate SRSF1 tends to bind on single-stranded RNA regions.

Predicted sequence and structural features of SRSF1 binding regions

To better compare RNAMoifModeler predictions with the SRSF1 binding motif reported previously, here we focus on octamer predictions. Consistent with the sequence consensus predicted by MEME [5], the reference sequence motif for SRSF1 proteins predicted using RNAMotifModeler is also 'AGAAGAAG' (Table 2 and Figure 3). Based on the predicted optimal parameters, we obtained an AUC of 0.875 (Figure 4(A)) and an maximal accuracy of 0.803 (Figure 4(B)), which are both higher than the MEME-based prediction, of which the AUC is 0.86 and maximal accuracy is 0.78 [5]. The optimal parameter matrix searched by RNAMotifModeler is presented in Table 2. The first row listed the reference sequence motif identified while the following four rows include retained binding affinity ratios caused by sequence alterations. To visualize the predicted SRSF1 sequence consensus more straightforwardly, positive gold-standard sequences were scanned to search binding sites with binding affinities higher than the threshold 0.138, based on which a sequence logo was created by Weblogo [19]. This motif is consistent with the positional weight matrix (PWM) identified by MEME using the same gold standard sequences in our previous study [5], and is similar to the motifs found by other groups [20-22]. The last row in Table 2 is constituted by unpaired probabilities for all nucleotides in the motif, indicating the optimal RNA secondary structure of SRSF1 binding regions. We note that every nucleotide of the predicted SRSF1 binding motif has a very high probability to be single-stranded, suggesting that SRSF1 proteins tend to bind on highly unpaired RNA regions.

Table 2 Optimal parameters of the octamer predicted by RNAMotifModeler.

	A	G	A	A	G	A	A	G
A	1.00	0.17	1.00	1.00	0.24	1.00	1.00	0.81
G	0.79	1.00	0.65	0.90	1.00	0.84	1.00	1.00
C	0.52	0.32	0.50	0.16	0.35	0.02	0.34	0.63
U	0.75	0.15	0.39	0.63	0.09	0.06	0.73	0.55
UP	0.99	0.96	0.99	0.99	0.98	0.99	0.92	0.83

The column headers represent the predicted reference motif. The first four rows indicate the retained binding affinity ratio corresponding to a sequence alteration from each nt in the reference motif. The last row shows the optimal unpaired probability (UP) for each nt of the predicted reference motif.

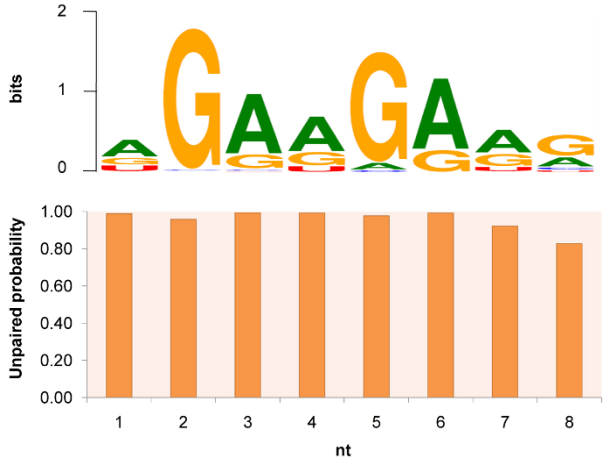

Figure 3 Sequence consensus logo (the upper panel) and unpaired probabilities (the lower panel) of SRSF1 predicted by RNAMotifModeler.

SRSF1-RNA binding sites are presented as single-stranded regions

Based on the predicted unpaired probabilities of nucleotides in the converged reference motif, we can conclude that SRSF1 tends to recognize purine-rich motifs in regions with a low degree of predicted secondary structure. To further test the hypothesis that RNA regions bound by SRSF1 proteins are significantly unpaired, we compared the unpaired probability of each nucleotide in the predicted binding sites with two different control sets of random'binding'sites. In our first control set, we randomly picked the same number of sites as the predicted binding sites in each positive gold standard sequence. P-values were computedfor each nucleotide based on Wilcoxon rank sum tests, with the alternative hypothesis that the unpaired probability in predicted binding sites are higher than controls. Indeed, all median unpaired probabilities of predicted binding sites are significantly higher than controls (Figure 5(A)). In the second control set, we first randomly selected the same number of exonic fragments that were not targeted by SRSF1 as the positive gold standards. The length of each exonic fragment equals its counterpart in the positive gold standard sequence. We subsequently drew the same number of random sites as the predicted binding sites from each control sequence. Wilcoxon tests were also performed between predicted binding sites and the second control set of random sites. Again, all the eight nucleotidesof binding sites in CLIP sequences are significantly more single-stranded (Figure 5(B)). The boxplots of the unpaired probabilities of the random sites in the

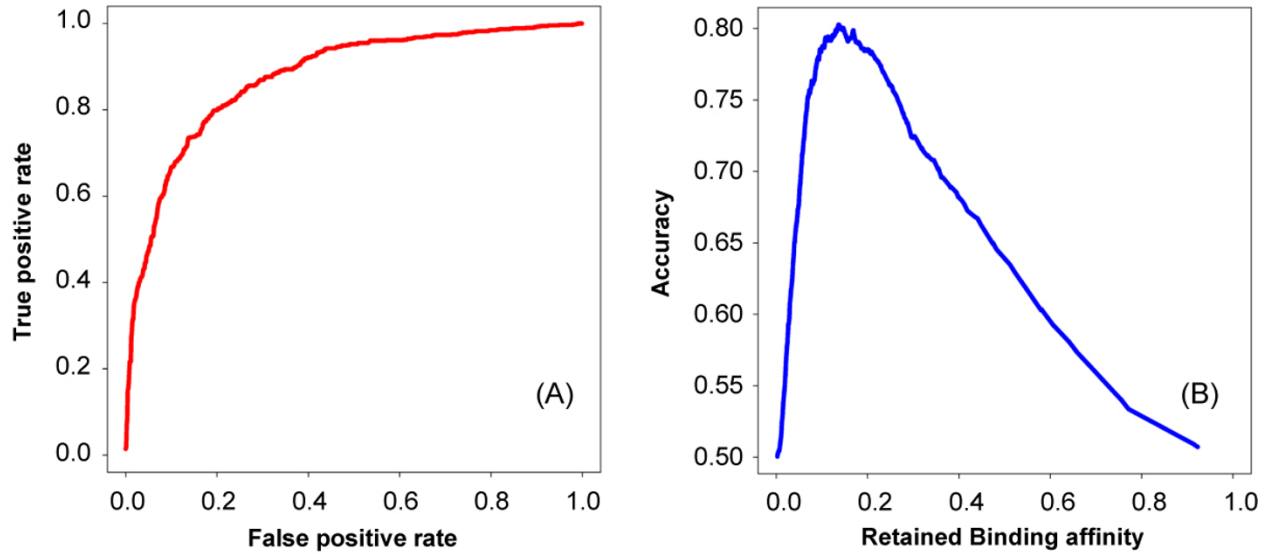

Figure 4 ROC curve and accuracy curve describing the prediction power of RNAMotifModeler for SRSF1 proteins.

two control sets and the predicted binding sites are shown in Figure 6(A), (B) and 6(C), respectively.

The two groups of Wilcoxon testsdemonstratethat binding sites predicted by RNAMotifModeler are not only more unpaired in positive gold standard sequences than those not targeted by SRSF1, but also less structured than by chance within themselves. More interestingly, comparing Figure 5(A) with Figure 3, we found that the second and fifth nt of SRSF1 motif have much stronger sequence specificities but lower *p*-values in the Wilcoxon tests against controls, while the third, fourth, sixth and seventh positions are more significantly single-stranded but have less sequence specificities, suggesting that both the sequence and secondary structure play complementary roles in the specificity of SRSF1-RNA interactions.

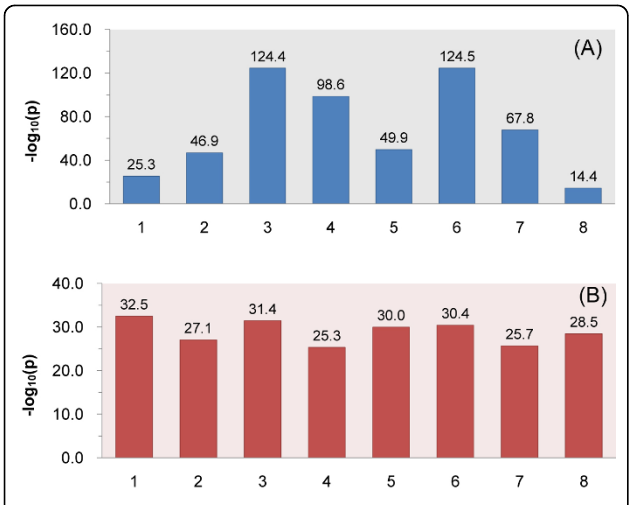

Figure 5 Results of Wilcoxon tests with the alternative hypothesis thatthe unpaired probabilities in the predicted binding sites are higher than negative controls randomly drawn from (A) the same positive gold-standard sequences, and (B) random sequences.

Comparisons of predictions before and after incorporating RNA structure information

RNAMotifModeler can be used to predict sequence consensus motifs enriched in CLIP data without evaluating the structural context of the co-purified RNA. We then investigated whether inclusion of the unpaired probablities for each nucleotide or not contributes to the consensus motif elucidated by RNAMotifModler. Using the same positive and negative gold-standard sequences, we identified the same reference motif 'AGAAGAAG' and a very similar sequence parameter matrix. However, we obtained anAUC of0.853 and a maximal accuracy of 0.789, suggesting a slightly reduced prediction power when discarding RNA secondary structure information (Additional file 2). Using the identified parameter matrix based on only sequenceswe predicted 2295 binding sites, of which 81% are commonly identified by incorporating RNA secondary structure information (Figure 7(A)). The unpaired probabilities of the other 437 binding sites are significantly lower than previously identified binding sites (Figure 7(B) and 7(C)). Except the third nucleotide of motif, all of the unpaired probabilities of these binding sites are even lower than background, indicating that binding sites predictions may result in a considerable number of false positives caused by ignoring RNA secondary structures. Bringing in RNA secondary structure information, we found 1046 more binding sites.

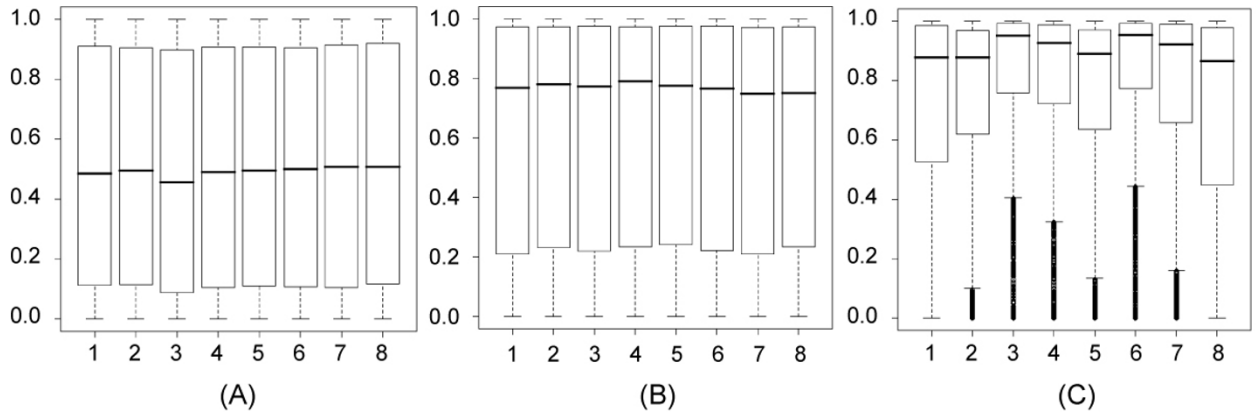

Figure 6 Unpaired probabilities of nucleotides of (A) a control set of binding sites randomly selected from negative goldstandard sequences (B) a control set of binding sites randomly selected from positive goldstandard sequences, and (C) binding sites predicted by RNAMotifModeler in positive gold standard sequences.

These binding sites may have low sequence specificities, but their binding affinities can be complemented by high structure specificities. Although the AUC increases only by 0.023 after introducing RNA secondary structure information, false positive and false negative binding sites are both significantly reduced.

Discussion

In recent years, there is an increasing interest in using high-throughput sequencing technology to study protein-RNA binding patterns, but almost all of current bioinformatic approaches used for this purpose do not take into account RNA secondary structures, which have been demonstrated to have critical impact on protein-RNA binding in previous biochemical experiments. Thus, the starting point of our proposed model–RNAMotifModeler is to predict both structural and sequence specificities of protein-RNA binding regions. We demonstrated the potential of RNAMotifModeler by an application to predicting binding specificities of SRSF1 proteins and obtained a reference motif of 'AGAAGAAG' with a parameter matrix including retained binding affinity ratios caused by sequence degeneracy, as well as probabilities for nucleotides being unpaired.

RNAMotifModeler incorporates RNA secondary structure using RNAfold-derived probabilities of nucleotides being paired with its neighbours. The preference for base-pairing probabilities over RNA secondary structures is due to a couple of concerns: a) It is very difficult to take into account RNA secondary structures directly in many real applications because of multiple RNA folding choices including optimal and sub-optimal structures; b) Unlike MEMERIS, RNAMotifModeler tries to identify the optimal structural feature that is expected to represent the base pairing probability for each nucleotidein motif. Therefore, we did not use measurements of single-strandedness of the entire protein-RNA binding regions in MEMERIS [12]. c) The base-pairing probabilities predicted by RNAfold program [18] encode all possible secondary structures.

We note from our prediction results that almost all unpaired probabilities of bases in the reference motif of SRSF1 predicted by RNAMotifModeler are close to 1, suggesting a very strong preference of SRSF1 to single-stranded RNA. The statistical significances were further proved by Wilcoxon tests on unpaired probabilities of nucleotides between predicted binding sites and randomly selected sites in random exonic fragments that were not targeted by SRSF1. Another group of Wilcoxon tests show that the unpaired probabilities of predicted binding sites are all significantly higher than thoserandomly selected in the same positive gold-standard sequences, indicating that SRSF1 proteins indeed have strong bias to single-stranded regions. These findings are consistent with previous conclusions in the literature. It is known that SRSF1 protein contains an arginine-serine rich region (RS domain) and two RNA recognition motifs (RRMs), through which SRSF1 recognizes specific RNA regions [23,24]. Importantly, RRM is one of single-stranded RNA-binding domains of proteins [25]. Comparing the sequence consensus and p-values derived from Wilcoxon tests, we propose that sequence and structural specificities may be two complementary factors that are both facilitating the binding site recognition of SRSF1.

RNAMotifModeler also provides an option to predict only sequence consensusmotif based on gold-standard protein-RNA binding sequences. In the specific application to SRSF1, we found that the prediction power is still comparable with MEME-based approach, although the AUC and maximum accuracy were both reduced when RNA secondary structure information was not

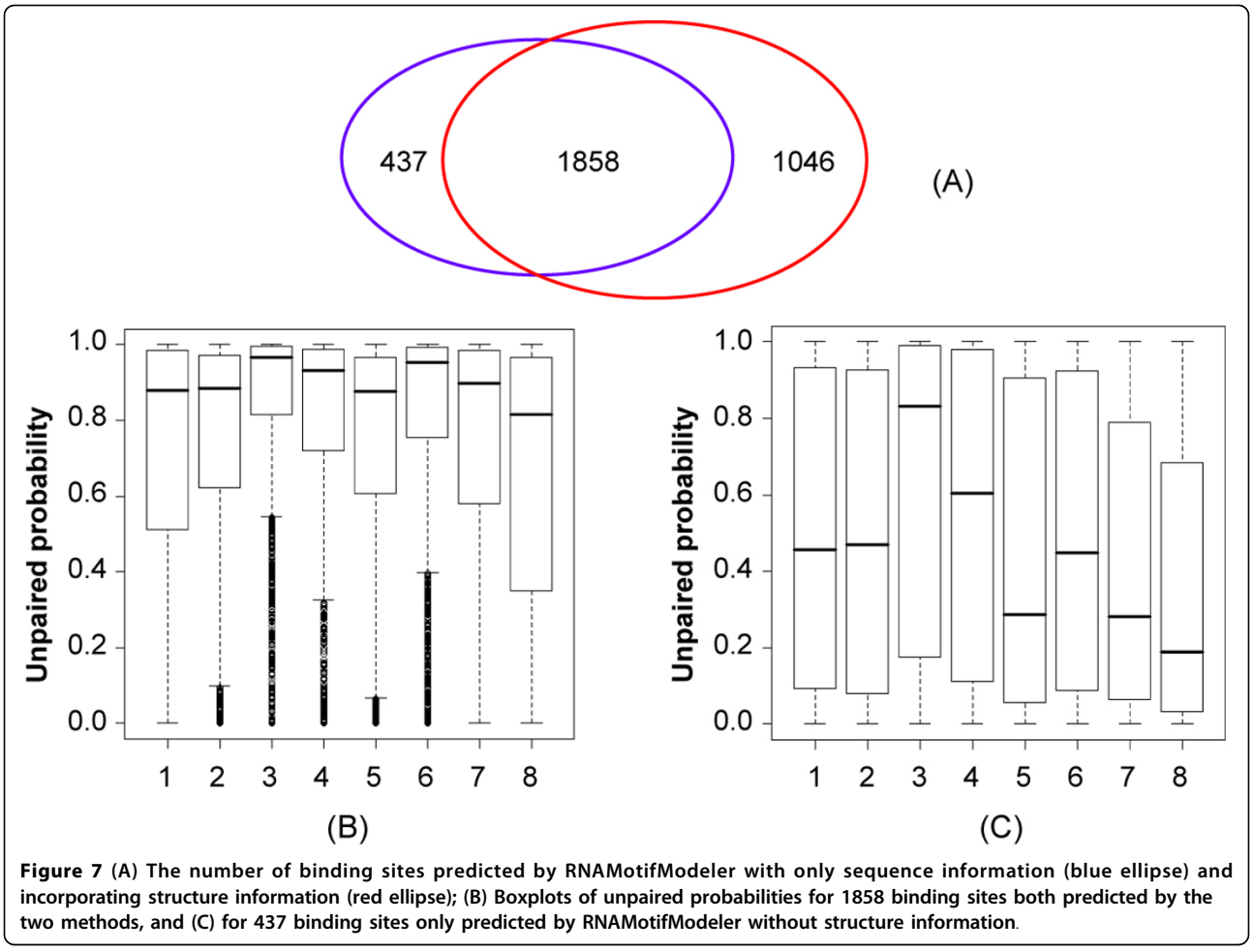

Figure 7 (A) The number of binding sites predicted by RNAMotifModeler with only sequence information (blue ellipse) and incorporating structure information (red ellipse); (B) Boxplots of unpaired probabilities for 1858 binding sites both predicted by the two methods, and (C) for 437 binding sites only predicted by RNAMotifModeler without structure information.

incorporated. Using predicted reference motif and sequence degeneracy parameters we identified 2295 binding sites, of which 437 are not included in the binding sites predicted after incorporating structure information. Interestingly, the unpaired probabilities of these 437 binding sites predicted by RNAfold are even lower than random sites selected by chance. Based on previous biological studies, we argue that these binding sites are probably false positives, although they satisfy the sequence specificity requirements.

Two parameters–the number of particles *M*and the contraction-expansion coefficient β of the Quantum Particle Swarm Optimization greatly affect the predicting accuracy of RNAMotifModeler. To estimate and set up these parameters prior to the optimization procedure, we did a series of hexamer motif searching tests with *M*enumerated from 10 to 10000 and β ranging from 0 to 1 for SRSF1 CLIP-seq data. The AUC scores resulted from optimizations using all parameter combinations are presented in 3D heatmaps (Additional file 3). We observed a much more rapid decrease in prediction power as β becomes lower when M is small. In contrast, when β is sufficiently high, the AUC score is not greatly affected by M. Thus, the greater M and β are, the higher prediction performance RNAMotifModeler would achieve. However, under the consideration of computational efficiency, we have to consider the time consumed in each test (Additional file 3). The time consumed is exponential to the increment of the number of particles, and is not actually controlled by β. When m is 100 and β equals 1.0, RNAMotifModeler achieved a high AUC score of 0.86 within three minutes. These two parameters are then selected for all the other optimizations for the SRSF1 dataset used in this study.

Convergence of optimization algorithms used in predicting protein-DNA or protein-RNA binding sites is a common concern due to a number of parameters to be fitted in model. In this report, we proposed motif searching pathways and motif searching graphs to inspect whether or not the algorithm of RNAMotifModeler indeed has a good convergence property. Importantly, the convergence of randomly initialized motif candidate to final targets turned out to be very fast. Thus, for short motifs, we suggest generate such a motif searching graph in order to

have a global landscape of all possible converged motifs and their possible relationships, as they are possibly parts of a common longer motif.

Despite our successful characterization of the binding features of SRSF1 proteins, our future work will be applying RNAMotifModeler to studying specificities of other RNA binding proteins such as fox2, NOVA and EWS, for which high-throughput sequences are available.

Methods

Predicting RNA base-pairing probabilities

One of the distinct features of RNAMotifModeler is that the information of secondary structures of the RNA regions bound by SRSF1 proteins is incorporated into the motif identification. For each nucleotide in the RNA fragment, we calculate the base pairing probability using the RNAfold function of the Vienna RNA package (version 1.7.1) [18]. The base pairing probability is used since it integrates likelihood of single-strandedness over multiple possible RNA secondary structures. For the CLIP-seq derived RNA fragments, these probabilities are generated based on the base pairing probability of base i being paired with base j, denoted as $p_{i,j}$. The binding probability of base i with all other neighbouring bases, defined as P_i, is calculated by:

$$P_i = \sum_{j=i+1}^{n_s} P_{ij} + \sum_{j=1}^{i-1} P_{ji}, \quad (1)$$

where n_s is the length of sequence s. Similar strategies are also used elsewhere [26,27].

Modelling protein-RNA binding affinities

In *RNAMotifModeler*, the consensus of each binding motif is defined by the following components: 1) The reference motif, a k-base RNA sequence on which the protein preferably binds; 2) Retained binding affinity despite of a one-nucleotide deviation from reference motif to the sequence of one binding sites. For each k-base motif, there are $3k$ retained binding affinities that describe all the possible deviations from reference motif. For instance, if the i-th base of the reference motif and a specific binding site is m_i and f_i, respectively, the retained binding affinity is defined as μ_{i,m_i,f_i}; 3) a vector that denotes the optimal base pairing probability of k bases in the motif $\boldsymbol{\theta} = (\theta_i)$; and 4) the penalty for the deviation from the optimal base pairing probability α. All these parameters will be optimized iteratively. A matching score describing the similarity between an RNA fragment (F) and a reference motif (R) is defined:

$$S_{R,F} = \max_{l=1}^{L-k+1} \left(S_{R,F,l}\right), \quad (2)$$

Where $S_{R,F,l}$ is the binding affinity for l-th binding site:

$$S_{R,F,l} = \prod_{i=1}^{k} \left(\left(\mu_{i,m_i,f_{l,i}}\right)\left(1 - \alpha \bullet \left|\theta_i - P_{f_{l,i}}\right|\right)\right), \quad (3)$$

where P_{f_i} represents the pairing probability of the i-th nucleotide in the RNA fragment F, calculated in Eq. (1). This matching score integrates the loss of binding affinity caused by both nucleotide and structure deviances from reference motif. We denote the parameter associated to the reference motif R as $\boldsymbol{\lambda_R} = (\boldsymbol{\mu},\boldsymbol{\theta}, \alpha)_{\boldsymbol{R}}$, where $\boldsymbol{\mu},\boldsymbol{\theta}$ and αrepresent the $3k$ retained binding affinities, optimal base pairing probability of k bases, and the penalty for the deviation from the optimal base pairing probability, respectively.

Identify the optimal reference motif from CLIP-seq data

We adopted an iterative approach to identify the optimal reference motif and its associated parameters, using a Quantum Particle Swarm Optimization algorithm (QPSO) [28]. The iterative strategy includes the selection of reference motif$R$, and optimization of the parameters associated to the reference motif $\boldsymbol{\lambda_R}$. The overall procedure includes the following steps:

1. Randomly select a motif candidate R_{init} from the motif searching space $\mathbf{M} = \{b_1b_2...b_k : b_1,b_2,...b_k \in$ {A,G, C,U}}as the reference motif.

2. Optimize the parameters for the reference motif by maximizing its ability for characterizing the CLIP-seq-derived RNA fragments.

Step 2.1. Parameter initiation. We first create M particles in the parameter space by randomly selecting numbers from U(0,1).

Step 2.2. Particle evaluation. For each particle (parameters), we evaluate its capability for distinguishing the CLIP-seq-derived RNA fragment from background sequences. We plot an ROC (Receiver Operating Characteristic) curve by adjusting the matching score threshold, calculated in Eq. (2). The quality of the parameter is evaluated based on the AUC (area under the curve) of the ROC plot.

Step 2.3. Particle update. Let $\boldsymbol{\lambda}_i^{selfbest}(t)$ and $\boldsymbol{\lambda}^{globalbest}(t)$ be the best individual particle and the population of particles has met at the t-th iteration. To guarantee convergence, each particle must converge to its local attractor $\boldsymbol{\lambda}_i^{pbest}$[28]. Compute $\boldsymbol{\lambda}_i^{pbest}(t)$ and the mean of the best positions of all particles $\boldsymbol{\lambda}^{mbest}(t)$ as follows:

$$\lambda_{i,j}^{pbest}(t) = (\varphi_1 \cdot \lambda_{i,j}^{selfbest}(t) + \varphi_2 \cdot \lambda_j^{globalbest}(t))/(\varphi_1 + \varphi_2) \quad (4)$$

$$\lambda_j^{mbest}(t) = \sum_{i=1}^{m} \lambda_{i,j}^{pbest}(t)/m, \quad (5)$$

where ϕ_1 and $\phi_2$are random variables following U(0,1);

QPSO employs Monte Carlo method to update parameters:

$$\lambda_{i,j}(t+1) = \begin{cases} \lambda_{i,j}^{pbest}(t) - \beta \cdot |\lambda_j^{mbest}(t) - \lambda_{i,j}(t)| \cdot \ln(1/u), & k \geq 0.5 \\ \lambda_{i,j}^{pbest}(t) + \beta \cdot |\lambda_j^{mbest}(t) - \lambda_{i,j}(t)| \cdot \ln(1/u), & k < 0.5 \end{cases}, \quad (6)$$

where β is called contraction-expansion coefficient controlling the convergence speed of QPSO; u and k are random variables which also follow U(0,1).

Repeat Step 2 and Step 3 until $|\boldsymbol{\lambda}^{globalbest}(t+1) - \boldsymbol{\lambda}^{globalbest}(t)| < \varepsilon$ repeatedly, in which ε is a tolerance used here as a criterion for the algorithm to terminate;

3. Based on the final parameter vector $\boldsymbol{\lambda}^{globalbest}$, the maximal binding affinity of motif candidate K in positive gold standard sequence F is:

$$a_{K,F} = \underset{\sigma \in \Omega_{K,F}}{\text{Max}}\ a_{K,F,\sigma}, \quad (7)$$

where $\Omega_{K,F}$ denotes the set of all binding sites for motif K in sequence F; $a_{K,F,\sigma}$ is also computed by Eq. (3).

Let n_s and n_m be the number positive gold standard sequences and the number of motif candidates, respectively. Let $S_{R_{init},F}$ be the maximal binding affinity computed using optimized parameters for the initial reference motif R_{init} in sequence F. Although R_{init} is a reference motif, $S_{R_{init},F}$ is not necessarily contributed by a binding site instance of R_{init}. In contrast, the 'real' reference motif contributes are always expected to contribute more to the binding affinities. Thus, to evaluate contributions of all motif candidatesto binding affinities ofpositive gold standard sequence, we define $\mathbf{c} = [c_{F,K}]_{F=1,2,\ldots,n_s,K=1,2,\ldots,n_m}$ as the motif contribution scorematrix:

$$c_{F,K} = \begin{cases} 0, & a_{K,F} \neq S_{R_{init},F} \\ 1, & a_{K,F} = S_{R_{init},F} \end{cases}, \quad (8)$$

and $\mathbf{v} = [v_K]_{K=1,2,\ldots,n_m}$ as the motif contribution score vector:

$$v_K = \sum_{F=1}^{n_s} c_{F,K}, \quad (9)$$

We denote the motif associated with the maximum score in $\mathbf{v}$ as R_{max}. If $R_{max} = R_{init}$, meaning the initialized reference motif accounts for the most contribution to the retained binding affinities, then we stop the iteration; otherwise, let R_{max} be the next R_{init}, and repeat steps 2 and 3 until convergence.

Motif searching pathways and graph

The route from original assumed reference motif to the final converged motif is called a *motif searching pathway*. Different initialized motifs may converge to different final motifs. Therefore, to investigate the convergence performance of RNAMotifModeler, it is important to enumerate all possible convergence pathways and find out what are the final converging points. For the specific example of SRSF1 protein, we used RNAMotifModeler to predict the optimal parameters for each one of 4096 motif candidates. All pathways are summarized and illustrated in a graph in Figure 2 using Cytoscape (version 2.6.1) [29]. Source motifs (all initial motifs), intermediary motifs (motifs which are neither final nor initial motifs) and destination motifs (converged motifs) are colored in blue, purple and red colors, respectively. The arrow from a source motif to an intermediary or destination motif denotes one motif transit. From the graph, we observe that the vast majority of original motifs transited to only three motifs, which we believe are the best candidates of reference motifs for SRSF1 proteins.

RBP binding motif logo

Although RNAMotifModeler provides a parameter matrix consisting of retained binding affinity ratios due to sequence mutations and structure alterations at each base, it is not straightforward for people to comprehend the sequence consensus. Thus, we provide an alternative way to generate a Positional Weight Matrix (PWM) and corresponding sequence logo. First of all, once RNAMotifModeler reaches a convergence, we obtain the optimal reference motif, estimated parameters, ROC curve, AUC score and the accuracy curve. At the peak of the accuracy curve we choosecorresponding binding affinity as a cutoff. Then, we trace back to each positive gold standard sequence and search all binding sites with binding affinities higher than the cutoff score. These predicted binding sitesare subsequently used to compute a corresponding PWM and create a sequence logo based on Weblogo [19].

Additional material

Additional file 1: Optimal 6nt and 7nt sequence-structural consensus for SRSF1 proteins predicted by RNAMotifModeler. The upper panel (A) and the lower panel (B) show the sequence and structural parameters identified for motif of length 6nt and 7nt, respectively.

Additional file 2: Prediction results based on RNAMotifModeler excluding the information of RNA secondary structure. (A) ROC curve (B) Accuracy curve, and (C) consensus sequence logo.

Additional file 3: 3D heatmaps illustrating (A) the prediction power and (B) time cost of RNAMotifModeler affected by the number of particles and the Contraction-Expansion coefficient which are two critical parameters of QPSO algorithm.

Acknowledgements

This work is supported by the grant from National Institutes of Health GM085121, CA113001, and AA017941 and the Indiana Genomics Initiative of Indiana University (supported in part by the Lilly Endowment, Inc.).

Author details
[1]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, IN 46202, USA. [2]College of Automation, Harbin Engineering University, Harbin, Heilongjiang 150001, China. [3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, IN 46202, USA. [4]Department of Molecular, Cellular and Developmental Biology, University of California Santa Cruz, Santa Cruz, California 95064, USA. [5]Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN 46202, USA.

Authors' contributions
XW and YL contributed to the design of the study. XW and YL designed and performed the computational modelling and drafted the manuscript. JRS provided the CLIP-seq data for SRSF1 proteins. XW, LJ, JRS, JL, KW and YL participated in coordination, discussions related to result interpretation and revision of the manuscript. All the authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 23 December 2011

References

1. Glisovic T, Bachorik JL, Yong J, Dreyfuss G: **RNA-binding proteins and post-transcriptional gene regulation.** *FEBS Lett* 2008, **582(14)**:1977-1986.
2. Buratti E, Muro AF, Giombi M, Gherbassi D, Iaconcig A, Baralle FE: **RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon.** *Mol Cell Biol* 2004, **24(3)**:1387.
3. Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR: **Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins.** *Nat Biotechnol* 2009, **27(7)**:667-670.
4. Schroeder R, Barta A, Semrad K: **Strategies for RNA folding and assembly.** *Nat Rev Mol Cell Biol* 2004, **5(11)**:908-919.
5. Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y: **Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts.** *Genome Res* 2009, **19(3)**:381-394.
6. Ule J, Jensen K, Mele A, Darnell RB: **CLIP: a method for identifying protein-RNA interaction sites in living cells.** *Methods* 2005, **37(4)**:376-386.
7. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: **CLIP identifies Nova-regulated RNA networks in the brain.** *Science* 2003, **302(5648)**:1212.
8. Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon YS, Zhang C, Yeo G, Black DL, Sun H: **Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping.** *Mol Cell* 2009, **36(6)**:996-1006.
9. Yeo G, Coufal N, Liang T, Peng G, Fu X, Gage F: **An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells.** *Nat Struct Mol Biol* 2009, **16(2)**:130-137.
10. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB: **An RNA map predicting Nova-dependent splicing regulation.** *Nature* 2006, **444(7119)**:580-586.
11. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *International Conference on Intelligent Systems for Molecular Biology (ISMB)* 1994, **2**:28-36.
12. Hiller M, Pudimat R, Busch A, Backofen R: **Using RNA secondary structures to guide sequence motif finding towards single-stranded regions.** *Nucleic Acids Res* 2006, **34(17)**:e117.
13. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q: **RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins.** *PLoS Comput Biol* **6(7)**:255-449.
14. Sanford JR, Coutinho P, Hackett JA, Wang X, Ranahan W, Caceres JF: **Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF.** *PLoS One* 2008, **3(10)**:e3369.
15. Tacke R, Manley JL: **The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities.** *EMBO J* 1995, **14(14)**:3540.
16. Sanford JR, Ellis J, Caceres JF: **Multiple roles of arginine/serine-rich splicing factors in RNA processing.** *Biochem Soc Trans* 2005, **33(3)**:443-446.
17. Karni R, De Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR: **The gene encoding the splicing factor SF2/ASF is a proto-oncogene.** *Nat Struct Mol Biol* 2007, **14(3)**:185-193.
18. Hofacker IL: **RNA secondary structure analysis using the Vienna RNA package.** *Curr Protoc Bioinformatics* 2004, **Chapter 12**:Unit 12.2.
19. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14(6)**:1188-1190.
20. Caputi M, Casari G, Guenzi S, Tagliabue R, Sidoli A, Melo CA, Baralle FE: **A novel bipartite splicing enhancer modulates the differential processing of the human fibronectin EDA exon.** *Nucleic Acids Res* 1994, **22(6)**:1018-1022.
21. Ramchatesingh J, Zahler AM, Neugebauer KM, Roth MB, Cooper TA: **A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer.** *Mol Cell Biol* 1995, **15(9)**:4898-4907.
22. Tacke R, Manley JL: **The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities.** *EMBO J* 1995, **14(14)**:3540-3551.
23. Ngo JCK, Giang K, Chakrabarti S, Ma CT, Huynh N, Hagopian JC, Dorrestein PC, Fu XD, Adams JA, Ghosh G: **A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1.** *Mol Cell* 2008, **29(5)**:563-576.
24. Hagopian JC, Ma CT, Meade BR, Albuquerque CP, Ngo JCK, Ghosh G, Jennings PA, Fu XD, Adams JA: **Adaptable molecular interactions guide phosphorylation of the SR protein ASF/SF2 by SRPK1.** *J Mol Biol* 2008, **382(4)**:894-909.
25. Auweter SD, Oberstrass FC, Allain FHT: **Sequence-specific binding of single-stranded RNA: is there a code for recognition?** *Nucleic Acids Res* 2006, **34(17)**:4943.
26. Hiller M, Pudimat R, Busch A, Backofen R: **Using RNA secondary structures to guide sequence motif finding towards single-stranded regions.** *Nucleic Acids Res* 2006, **34(17)**:e117.
27. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29(6-7)**:1105-1119.
28. Sun J, Feng B, Xu W: **Particle swarm optimization with particles having quantum behavior.** *IEEE 2004* 2004, 325-331.
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13(11)**:2498.

doi:10.1186/1471-2164-12-S5-S8
Cite this article as: Wang *et al.*: Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1. *BMC Genomics* 2011 **12**(Suppl 5):S8.